Multivariate and Semiparametric Kernel Regression*

Wolfgang HÄRDLE
Marlene MÜLLER

Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin, Germany

March 11, 1997

The paper gives an introduction to theory and application of multivariate and semiparametric kernel smoothing. Multivariate nonparametric density estimation is an often used pilot tool for examining the structure of data. Regression smoothing helps in investigating the association between covariates and responses. We concentrate on kernel smoothing using local polynomial fitting which includes the Nadaraya–Watson estimator. Some theory on the asymptotic behavior and bandwidth selection is provided. In the applications of the kernel technique, we focus on the semiparametric paradigm. In more detail we describe the single index model (SIM) and the generalized partial linear model (GPLM).

*To appear in: M.G. Schimek (Ed.), *Smoothing and Regression. Approaches, Computation and Application*, 1996.

The research for this paper was supported by Sonderforschungsbereich 373 at the Humboldt-University Berlin. The work of M. Müller was supported in part by CentER, Tilburg University (The Netherlands). The paper is printed using funds made available by the Deutsche Forschungsgemeinschaft.

Contents

Nonparametric smoothing methods serve three essential needs in statistical data analysis. First they provide a flexible analysis tool, often based on interactive graphical data representation (Scott, 1992). Second they help in constructing a model from observations, for example by comparison with concurrent models (Müller, 1988). Third they provide pilot estimators in adaptation problems, see Newey and Stoker (1993). Here we present the multivariate kernel smoother, examine the asymptotic properties of both density and regression estimators, and review applications of this technique in semiparametric statistics.

1 Multidimensional Smoothing with Kernels

In this section we review kernel smoothing methods for density and regression function estimation. Many ideas, in particular for asymptotics, bandwidth choice and graphical representation, are similar for both purposes. We can however only introduce a small part on the available material. In particular, for the regression case we restrict the presentation on the random design case. For a more detailed presentation of the subject we refer to the monographs by Härdle (1990; 1991), Scott (1992), Wand and Jones (1995) and Fan and Gijbels (1995).

1.1 Multivariate Kernel Density Estimation

The goal of multivariate nonparametric density estimation is to approximate the probability density function (pdf) $f(t) = f(t_1, \ldots, t_q)$ of the random variables $T = (T_1, \ldots, T_q)^T$. The multivariate kernel density estimator in the q-dimensional case is defined as

$$\widehat{f}_h(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1 \ldots h_q} \mathcal{K} \left(\frac{T_{i1} - t_1}{h_1}, \ldots, \frac{T_{iq} - t_q}{h_q} \right), \tag{1}$$

$\mathcal{K}$ denoting a multivariate kernel function $\mathcal{K} : I\!R^q \to I\!R$. Note, that (1) assumes that the bandwidth h is a vector of bandwidths $h = (h_1, \ldots, h_q)^T$.

What form shall the multidimensional kernel function $\mathcal{K}(u) = \mathcal{K}(u_1, \ldots, u_q)$ take on? The easiest solution is to use a *multiplicative* kernel

$$\mathcal{K}(u) = K(u_1) \cdot \ldots \cdot K(u_q) \tag{2}$$

with K denoting an univariate kernel function. For univariate kernels with support $[-1, 1]$ (as the Epanechnikov kernel $K(u) = 0.75(1-u^2)\,\mathrm{I}(|u| \leq 1)$) observations in a cube around t are used to estimate the density at the point t. An alternative is to use a genuine multivariate kernel function $\mathcal{K}(u)$, as e.g. the multivariate Epanechnikov

$$\mathcal{K}(u) \propto (1 - u^T u)\ \mathrm{I}(u^T u \leq 1).$$

This type of multivariate kernels can be obtained from univariate by defining

$$\mathcal{K}(u) \propto K(\|u\|), \tag{3}$$

where $\|u\| = \sqrt{u^T u}$ denotes the Euclidean norm of the vector u. Note that we use $\propto$ to indicate that the appropriate constant has to be multiplied. Kernels of the form (3) use observations from a ball around t to estimate the pdf at t. This type of kernels is usually called *spherical* or *radialsymmetric* since $\mathcal{K}(u)$ has the same value for all u on a sphere around zero. Figure 1 shows the contour lines from a bivariate product and a bivariate radialsymmetric kernel on the left and right hand side, respectively.

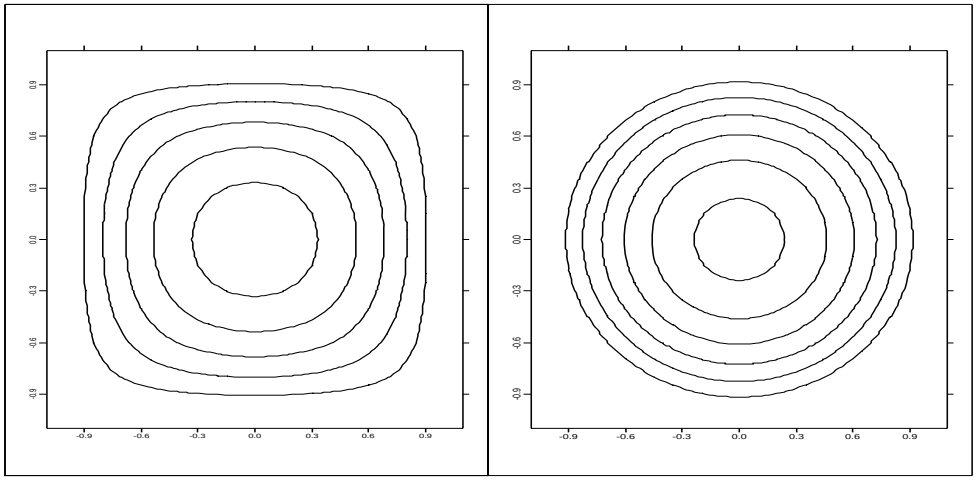


Figure 1: Contours from bivariate product (left) and bivariate radialsymmetric (right) Epanechnikov kernel.

Note that the kernel weights in Figure 1 correspond to equal bandwidth in each direction, i.e. $h = (h_1, h_2)^T = (1, 1)^T$. When we use different bandwidths, the observations around t in the density estimate $\widehat{f_h}(x)$ will be used with different weights in both dimensions.

Another approach is to use a nonsingular, symmetric *bandwidth matrix* $\mathbf{H}$. The general form for the multivariate density estimator is then

$$\widehat{f}_{\mathbf{H}}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\det(\mathbf{H})} \mathcal{K}\left\{\mathbf{H}^{-1}(T_i - t)\right\} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}\left(T_i - t\right), \tag{4}$$

see Silverman (1986) and Scott (1992). Here we introduce the short notation

$$\mathcal{K}_{\mathbf{H}}(\bullet) = \frac{1}{\det(\mathbf{H})} \mathcal{K}(\mathbf{H}^{-1}\bullet)$$

analogously to K_h in the one–dimensional case. A bandwidth matrix includes all simpler cases as special cases. An equal bandwidth h in all dimensions as in (1) corresponds to

$\mathbf{H} = h\mathbf{I}_q$ where $\mathbf{I}_q$ denotes the $q \times q$ identity matrix. Different bandwidths as in (1) are equivalent to $\mathbf{H} = \mathrm{diag}(h_1, \ldots, h_q)$, the diagonal matrix with elements $h_1, \ldots, h_q$.

What effect has the inclusion of off–diagonal elements? We will see that a good rule of thumb is to use a bandwidth matrix proportional to $\widehat{\mathbf{\Sigma}}^{-1/2}$ where $\widehat{\mathbf{\Sigma}}$ is the covariance matrix of the data. Hence, using such a bandwidth corresponds to a transformation of the data, so that they have an identity covariance matrix. As a consequence we can use bandwidth matrices to correct for correlation between the components of T. We have plotted the contour curves of product and radialsymmetric Epanechnikov weights with bandwidth matrix

$$\mathbf{H} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}^{1/2},$$

i.e. $\mathcal{K}_{\mathbf{H}}(u) = \mathcal{K}(\mathbf{H}^{-1}u)/\det(\mathbf{H})$, in Figure 2.

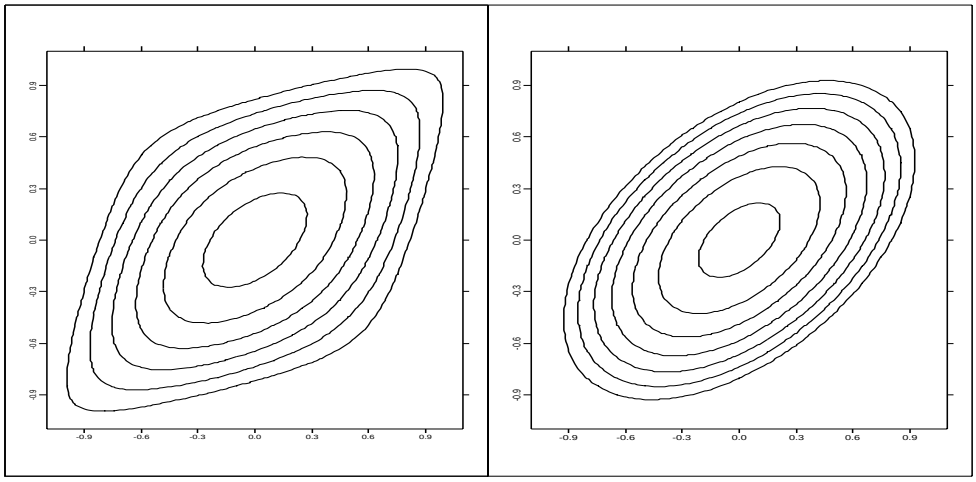


Figure 2: Contours from bivariate product (left) and bivariate radialsymmetric (right) Epanechnikov kernel. Bandwidth matrix.

In the following we will consider statistical properties as bias, variance, the issue of bandwidth selection and applications for this estimator. We formulate all results for estimators with bandwidth matrices and multivariate kernel function $\mathcal{K}$.

1.1.1 Bias, Variance and Asymptotics

A consequence of the standard assumption on the non–negative kernel $\mathcal{K}$

$$\int \mathcal{K}(u)\, du = 1 \tag{5}$$

is that the estimate $\widehat{f}_{\mathbf{H}}$ is a density function, i.e. $\int \widehat{f}_{\mathbf{H}}(t)\, dt = 1$. The estimate is consistent in any point t of continuity of f:

$$\widehat{f}_{\mathbf{H}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}\left(T_i - t\right) = f(t) + o_p(1), \tag{6}$$

if $n \to \infty$, $\mathbf{H} \to 0$ and $n\mathbf{H} \to \infty$, see e.g. Ruppert and Wand (1994). The derivation of the mean squared error *MSE* and the mean integrated squared error *MISE* is analogous to the one–dimensional case. We will sketch the asymptotic expansions and concentrate on the asymptotic mean integrated squared error *AMISE*.

As usual, *AMISE* has a bias part *AIB* and a variance part *AIV*. The bias of $\widehat{f}_{\mathbf{H}}(t)$ is $E\,\widehat{f}_{\mathbf{H}}(t) - f(t)$ and the integrated squared bias is

$$IB(\mathbf{H}) = \int \{E\,\widehat{f}_{\mathbf{H}}(t) - f(t)\}^2\,dt.$$

The asymptotic integrated squared bias $AIB(\mathbf{H})$ is the first order term of $IB(\mathbf{H})$, i.e.

$$\frac{IB(\mathbf{H}) - AIB(\mathbf{H})}{AIB(\mathbf{H})} = o(1)$$

as $\mathbf{H} \to 0$, $n \to \infty$ and $n\mathbf{H} \to \infty$. Define now the integrated variance

$$IV(\mathbf{H}) = \int E\{\widehat{f}_{\mathbf{H}}(t) - E\,\widehat{f}_{\mathbf{H}}(t)\}^2\,dt$$

and the asymptotic integrated variance *IV* accordingly to *IB*. Then the asymptotic mean integrated squared error *AMISE* can be calculated as

$$AMISE(\mathbf{H}) = AIB(\mathbf{H}) + AIV(\mathbf{H}). \tag{7}$$

A detailed derivation of the components of *AMISE* can be found in Scott (1992) or Wand and Jones (1995) and the references therein. As in the univariate case we use a second order Taylor expansion. Here and in the following we denote with ∇_f the gradient and with $\mathcal{H}_f$ the Hessian matrix of second order partial derivatives of a function (here f). Then the Taylor expansion of $f(\bullet)$ around t is

$$f(t+u) = f(t) + u^T\nabla_f(t) + \frac{1}{2}u^T\mathcal{H}_f(t)t + o(u^Tu),$$

see Wand and Jones (1995, p. 94). This leads to the expression

$$\begin{aligned} E\,\widehat{f}_{\mathbf{H}}(t) &= \int \mathcal{K}_{\mathbf{H}}(u-t)\,f(u)\,du = \int \mathcal{K}(s)\,f(t+\mathbf{H}s)\,ds \\ &\approx \int \mathcal{K}(s)\,\left\{f(t) + s^T\mathbf{H}^T\nabla_f(t) + \frac{1}{2}s^T\mathbf{H}^T\mathcal{H}_f(t)\mathbf{H}s\right\}\,ds. \end{aligned} \tag{8}$$

If we assume additionally to (5)

$$\int u\mathcal{K}(u)\,du = 0_q, \tag{9}$$

$$\int uu^T\mathcal{K}(u)\,du = \mu_2(\mathcal{K})\mathbf{I}_q, \tag{10}$$

then (8) yields $E\,\widehat{f}_{\mathbf{H}}(t) - f(t) \approx \frac{1}{2}\mu_2(\mathcal{K})\,\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_f(t)\mathbf{H}\}$, hence

$$AIB(\mathbf{H}) = \frac{1}{4}\mu_2^2(\mathcal{K})\int \left[\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_f(t)\mathbf{H}\}\right]^2\,dt. \tag{11}$$

As in univariate density estimation, the leading term of the variance part is the second moment of the estimate, i.e.

$$\begin{aligned} Var\left\{\widehat{f}_{\mathbf{H}}(t)\right\} &= \frac{1}{n}\int \left\{\mathcal{K}_{\mathbf{H}}(u-t)\right\}^2 du - \frac{1}{n}\left\{E\,\widehat{f}_{\mathbf{H}}(t)\right\}^2 \\ &\approx \int \frac{1}{n\det(\mathbf{H})}\,\mathcal{K}^2(s)\,f(t+\mathbf{H}s)\,ds \\ &\approx \int \frac{1}{n\det(\mathbf{H})}\,\mathcal{K}^2(s)\,\left\{f(t)+s^T\mathbf{H}^T\nabla_f(t)\right\}\,ds \\ &\approx \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2\,f(t), \end{aligned} \tag{12}$$

with $\|\mathcal{K}\|_2$ denoting the q-dimensional L_2-norm of $\mathcal{K}$. Hence

$$AIV(\mathbf{H}) = \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2 \tag{13}$$

and in summary we get the following *AMISE* formula for the multivariate kernel density estimator

$$AMISE(\mathbf{H}) = \frac{1}{4}\mu_2^2(\mathcal{K})\int\left[\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_f(t)\mathbf{H}\}\right]^2\,dt + \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2. \tag{14}$$

Let us now turn to the problem how to choose the *AMISE* optimal bandwidth. Again this is the bandwidth which balances bias–variance tradeoff in *AMISE*. Denote h a scalar, such that $\mathbf{H} = h\mathbf{H}_0$ and $\det(\mathbf{H}_0) = 1$. Then *AMISE* can be written as

$$AMISE(\mathbf{H}) = \frac{1}{4}\,h^4\,\mu_2^2(\mathcal{K})\int\left[\mathrm{tr}\{\mathbf{H}_0^T\mathcal{H}_f(t)\mathbf{H}_0\}\right]^2\,dt + \frac{1}{nh^q}\|\mathcal{K}\|_2^2.$$

If we only allow changes in h the optimal orders for the smoothing parameter h and *AMISE* are

$$h_0 = O(n^{-1/(4+q)}), \quad AMISE(h_0\mathbf{H}_0) = O(n^{-4/(4+q)}).$$

Hence, this density estimator has a rather slow rate of convergence, especially if q is large. If we consider $\mathbf{H} = h\mathbf{I}_q$ (the same bandwidth in all q dimensions) and we fix the sample size n, then the *AMISE* optimal bandwidth has to be considerably larger than in the one–dimensional case to make sure that the estimate has reasonably small variability. Some ideas of comparable sample sizes to reach the same quality of the density estimates over different dimensions can be found in Silverman (1986, p. 94) and Scott and Wand (1991). Moreover, the computational effort of this technique increases with the number of dimensions q. Therefore, multidimensional density estimation is usually not practically applied if $q \geq 5$.

1.1.2 Bandwidth selection and Graphical Representation

The problem of an automatic, data-driven choice of the bandwidth $\mathbf{H}$ is of great importance in the multivariate case. In one or two dimensions we may choose an "appropriate" bandwidth interactively by looking at the sequence of density estimates for different bandwidths. But how can this be done in three, four or more dimensions? The problem of graphical representation arises, which we address next.

Theoretically the bandwidth selection problem can be handled as in the one–dimensional case. Typically, one searches for a global bandwidth $\mathbf{H}$ or a local bandwidth $\mathbf{H}(t)$. Two approaches are frequently used in both cases

- plug–in bandwidths, in particular "rule–of–thumb" bandwidths,
- resampling methods, in particular cross–validation and bootstrap.

We will introduce generalizations for Silverman's rule–of–thumb and least squares cross–validation to stress the analogy with the one–dimensional bandwidth selectors.

Rule–of–thumb Bandwidth Rule–of–thumb bandwidth selection provides a formula arising from a reference distribution. Obviously, the pdf of a multivariate normal distribution $N_q(\mu, \mathbf{\Sigma})$ is a good candidate for a reference distribution in the multivariate case. Suppose that the kernel $\mathcal{K}$ is Gaussian, i.e. the pdf of $N_q(0_q, \mathbf{I}_q)$. Note that $\mu_2(\mathcal{K}) = 1$ and $\|\mathcal{K}\|_2^2 = 2^{-q}\pi^{-q/2}$ in this case. Hence, from (14) and the fact that

$$\int [\mathrm{tr}\{\mathbf{H}^T \mathcal{H}_f(t) \mathbf{H}\}]^2 \, dt = \frac{1}{2^{q+2}\pi^{q/2} \det(\mathbf{\Sigma})^{1/2}} \left[2\,\mathrm{tr}(\mathbf{H}^T \mathbf{\Sigma}^{-1} \mathbf{H})^2 + \{\mathrm{tr}(\mathbf{H}^T \mathbf{\Sigma}^{-1} \mathbf{H})\}^2\right]$$

we can easily derive rule–of–thumb formulae for different assumptions on $\mathbf{H}$ and $\mathbf{\Sigma}$.

In the simplest case, i.e. that we consider $\mathbf{H}$ and $\mathbf{\Sigma}$ to be diagonal matrices $\mathbf{H} = \mathrm{diag}(h_1, \ldots, h_q)$ and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$, this leads to

$$\widetilde{h}_j = \left(\frac{4}{q+2}\right)^{1/(q+4)} n^{-1/(q+4)} \sigma_j. \tag{15}$$

Note that this formula coincides with Silverman's rule–of–thumb in the case $q = 1$, see Silverman (1986, p. 45). Replacing the σ_j's by estimates and noting the first factor is always between 0.924 and 1.059, we arrive at Scott's rule

$$\widehat{h}_j = n^{-1/(q+4)} \widehat{\sigma}_j, \tag{16}$$

see Scott (1992, p. 152).

It is difficult to derive the rule–of–thumb for general $\mathbf{H}$ and $\mathbf{\Sigma}$. However, (15) shows that it might be a good idea to choose the bandwidth matrix $\mathbf{H}$ proportional to $\mathbf{\Sigma}^{1/2}$. In this case we get as generalization of Scott's rule

$$\widehat{\mathbf{H}} = n^{-1/(q+4)} \widehat{\mathbf{\Sigma}}^{1/2}. \tag{17}$$

We remark that this rule is equivalent to apply a Mahalanobis transformation on the data (to transform the estimated covariance matrix to identity), then to compute the kernel estimate with equal bandwidths $h = n^{1/(q+4)}$ and finally to retransform the estimated pdf back to the original scale.

But before we go on with applications, let us consider what we can do, if we want to use a kernel different from the Gaussian. The idea of canonical kernels by Marron and Nolan (1988) can be easily extended to the multivariate case. Consider a kernel $\mathcal{K}$ and all equivalent kernel functions $\mathcal{K}_\delta = \delta^{-1}\mathcal{K}(\bullet/\delta)$ with $\delta \geq 0$. Although that δ is a scalar, it is working on q–variates arguments of $\mathcal{K}$. Now we have $\|\mathcal{K}_\delta\|_2^2 = \delta^{-q}\|\mathcal{K}\|_2^2$ and $\mu_2(\mathcal{K}_\delta) = \delta^2\mu_2(\mathcal{K})$. As in the one–dimensional case we choose δ such that the bias-variance tradeoff in $AMISE(\mathbf{H}, \mathcal{K}_\delta)$ is independent of $\mathcal{K}_\delta$. This yields

$$\mu_2^2(\mathcal{K}_{\delta_0}) = \|\mathcal{K}_{\delta_0}\|_2^2 \iff \delta_0 = \left\{ \frac{\|\mathcal{K}\|_2^2}{\mu_2^2(\mathcal{K})} \right\}^{1/(q+4)}.$$

δ_0 again is called *canonical bandwidth* of the kernel $\mathcal{K}$. Denote now $\mathcal{K}^A$ a kernel function with canonical bandwidth δ_0^A and $\mathcal{K}^B$ a kernel function with canonical bandwidth δ_0^B. Suppose we have used $\mathbf{H}_A$ with kernel $\mathcal{K}^A$ and we want to recompute the kernel density estimate with kernel $\mathcal{K}^B$. Then it holds

$$AMISE(\mathbf{H}_A, \mathcal{K}^A) \approx AMISE(\mathbf{H}_B, \mathcal{K}^B)$$

if

$$\mathbf{H}_B = \frac{\delta_0^B}{\delta_0^A} \mathbf{H}_A, \tag{18}$$

which allows to adjust bandwidths for different kernel as in the one–dimensional case.

Let us consider an example. Suppose we want to use the product Quartic kernel $\mathcal{K}^Q$ instead of the q-dimensional Gaussian $\mathcal{K}^G$ which is faster in direct computation because of its compact support on $[-1, 1]$. Which is the equivalent rule–of–thumb to (17) in this case? Here we have $\delta_0^G = \{1/(2\sqrt{\pi})\}^{q/(q+4)}$ and $\delta_0^Q = (49 \cdot 5^q/7^q)^{1/(q+4)}$ which gives the canonical bandwidths in Table 1 for dimensions $q = 1, \ldots, 5$.

The fourth column of Table 1 gives the factor which the rule–of–thumb bandwidth matrix in (17) needs to be multiplied with to obtain the rule–of–thumb bandwidth for the multiplicative Quartic kernel. Of course all rule–of–thumb bandwidths for other kernel functions can be calculated in a similar way.

q	δ_0^G	δ_0^Q	δ_0^Q/δ_0^G
1	0.7764	2.0362	2.6226
2	0.6558	1.7100	2.6073
3	0.5814	1.5095	2.5964
4	0.5311	1.3747	2.5883
5	0.4951	1.2783	2.5820

Table 1: Bandwidth adjusting factors for Gaussian and multiplicative Quartic Kernel for different dimensions q.

For a product kernel $\mathcal{K}$ holds $\mu_2(\mathcal{K}) = \mu_2(K)$ and $\|\mathcal{K}\|_2 = \|K\|_2^q$ when K denotes the corresponding univariate kernel. A table of values $\mu_2(K)$, $\|K\|_2^2$ can be found in Härdle (1991, p.239) for example.

Principally, all plug–in methods for the one–dimensional kernel density estimation can be extended to the multivariate case. See Wand and Jones (1994) for details on multivariate plug–in bandwidth selection.

Cross–validation As we mentioned before, the cross–validation method is fairly independent of the special structure of the parameter or function estimate. Considering the bandwidth choice problem, cross–validation techniques allow to adapt to a wider class of density functions f than the rule–of–thumb approach. (Remember that the rule–of–thumb bandwidth is optimal for the reference pdf, hence it may fail for multimodal densities for instance.)

Recall, that in contrast to the rule–of–thumb approach, least squares cross–validation for density estimation aims to estimate the *ISE* optimal bandwidth. Here we approximate the integrated squared error

$$\begin{aligned} ISE(\mathbf{H}) &= \int \{\widehat{f}_{\mathbf{H}}(t) - f(t)\}^2 \, dt \\ &= \int \widehat{f}_{\mathbf{H}}^2(t) \, dt - 2 \int \widehat{f}_{\mathbf{H}}(t) f(t) \, dt + \int f^2(t) \, dt. \end{aligned} \tag{19}$$

Apparently, this is the same formula as in the the one–dimensional case and with the same arguments the last term of (19) can be ignored. The first term again can be easily calculated from the data. Hence, only the second term of (19) is unknown and has to be estimated. However, observe that $\int \widehat{f}_{\mathbf{H}}(t) f(t) \, dt = E \, \widehat{f}_{\mathbf{H}}(T)$, where the only new aspect now is that T is q–dimensional. As in the one–dimensional case we estimate this term by a leave–one–out estimator

$$E \, \widehat{\widehat{f}_{\mathbf{H}}}(T) = \frac{1}{n} \sum_{i=1}^{n} \widehat{f}_{\mathbf{H},-i}(T_i)$$

where

$$\widehat{f}_{\mathbf{H},-i}(t) = \frac{1}{n-1} \sum_{i \neq j, j=1}^{n} \mathcal{K}_{\mathbf{H}}(T_j - t).$$

This yields the multivariate cross–validation criterion as a straightforward generalization of CV in the one–dimensional case:

$$CV(\mathbf{H}) = \frac{1}{n^2 \det(\mathbf{H})} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{K} \star \mathcal{K} \left\{ \mathbf{H}^{-1}(T_j - t_i) \right\} - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathcal{K}_{\mathbf{H}}(T_j - T_i).$$

The difficulty comes in by the fact that the bandwidth is now a $q \times q$ matrix $\mathbf{H}$. In the most general case, this means, we have to minimize over $q(q+1)/2$ parameters. Still, if we assume $\mathbf{H}$ to be a diagonal matrix, this remains a q–dimensional optimization problem. This holds as well for other cross–validation approaches. Multivariate resampling methods for bandwidth selection are discussed in more detail in Sain, Baggerly and Scott (1994).

Graphical Representation Consider now the problem to graphically display a multivariate density estimate. Assume first $q = 2$. Here we are still able to show the density estimate in a 3-dimensional plot. This is in particular useful if the estimated function can be rotated on the computer screen interactively. For a two-dimensional presentation a contour plot gives often more insight to the structure of the data.

In the following, we will use the credit data from Fahrmeir and Hamerle (1984), Fahrmeir and Tutz (1994) for illustration. This data set consists of $n = 1000$ clients, 700 paid a credit back without problems, 300 did not. Among a number of categorical variables (running account, previous credits, purpose, personal attributes etc.) three continuous variables are available: duration and amount of credit as well as age.

Figures 3, 4 (upper panels) display a two-dimensional density estimate

$$\widehat{f}_h(t) = \widehat{f}_h(t_1, t_2)$$

for log(duration, log(amount and log(amount, log(age, respectively. We use the subscript h to indicate that we used a diagonal bandwidth matrix $\mathbf{H} = \mathrm{diag}(h_1, h_2)$.

Additionally, Figures 3, 4 (lower panels) gives contour plots of these density estimates. It is easily observed, that both distributions are rather symmetric. This is due to the logarithmic transformation. In the duration direction a typical bimodal structure can be recognized. This slightly reproduces in the amount direction. Obviously, both variables are related with positive correlation.

Here, the bandwidth was chosen accordingly to the general "rule–of–thumb" (17), which tends to oversmooth multimodal structures of the data. In fact, the durations of credits are multiples of 6 months in most case. The two clear modes that we observe are those for durations 12 and 24 months. In all applications of this paper we use the

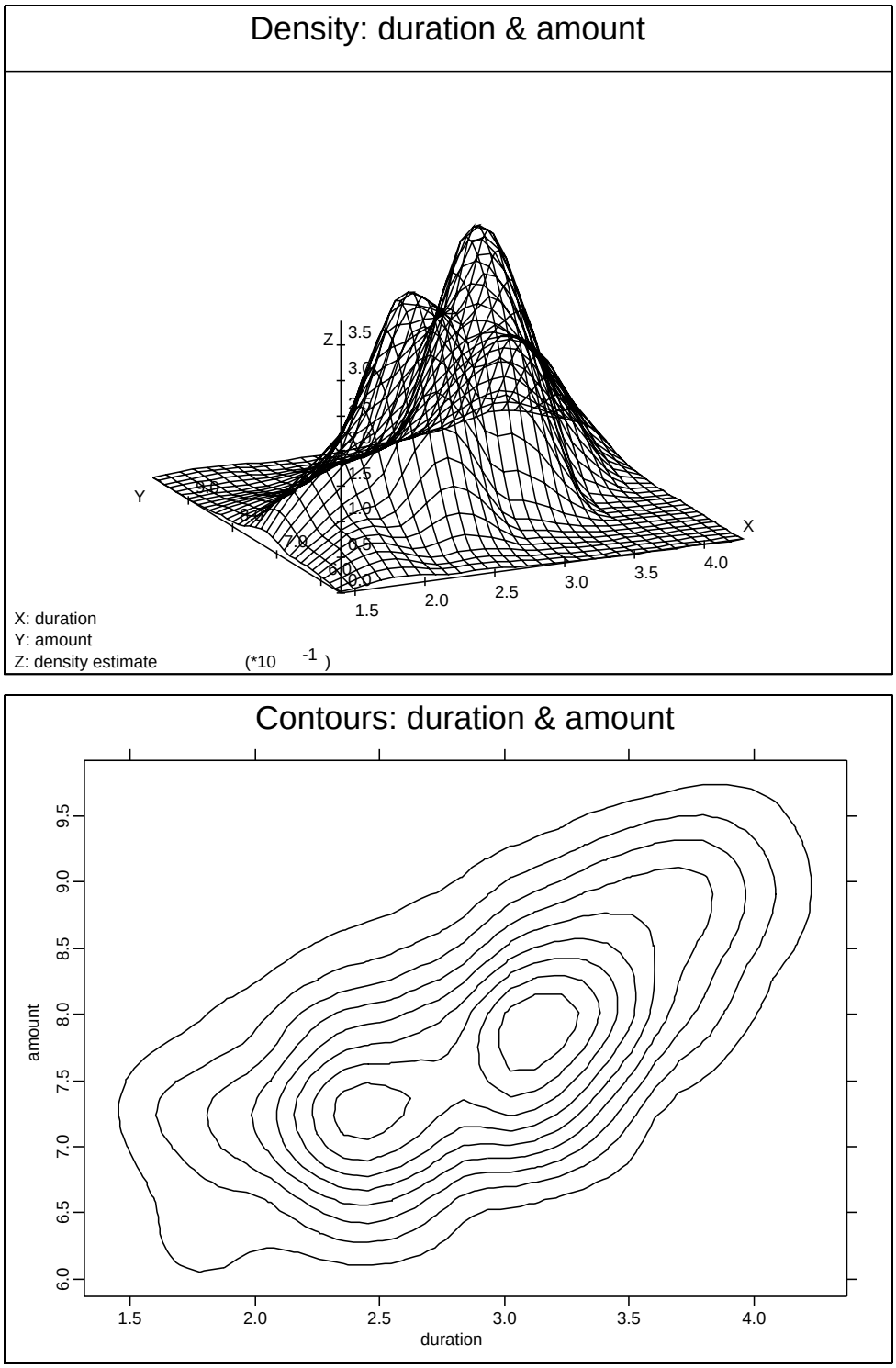


Figure 3: Two–dimensional density estimate (upper panel) and density contours (lower panel) for duration and amount. $h_1 = 0.48$, $h_2 = 0.64$. Credit data, Fahrmeir and Hamerle (1984).

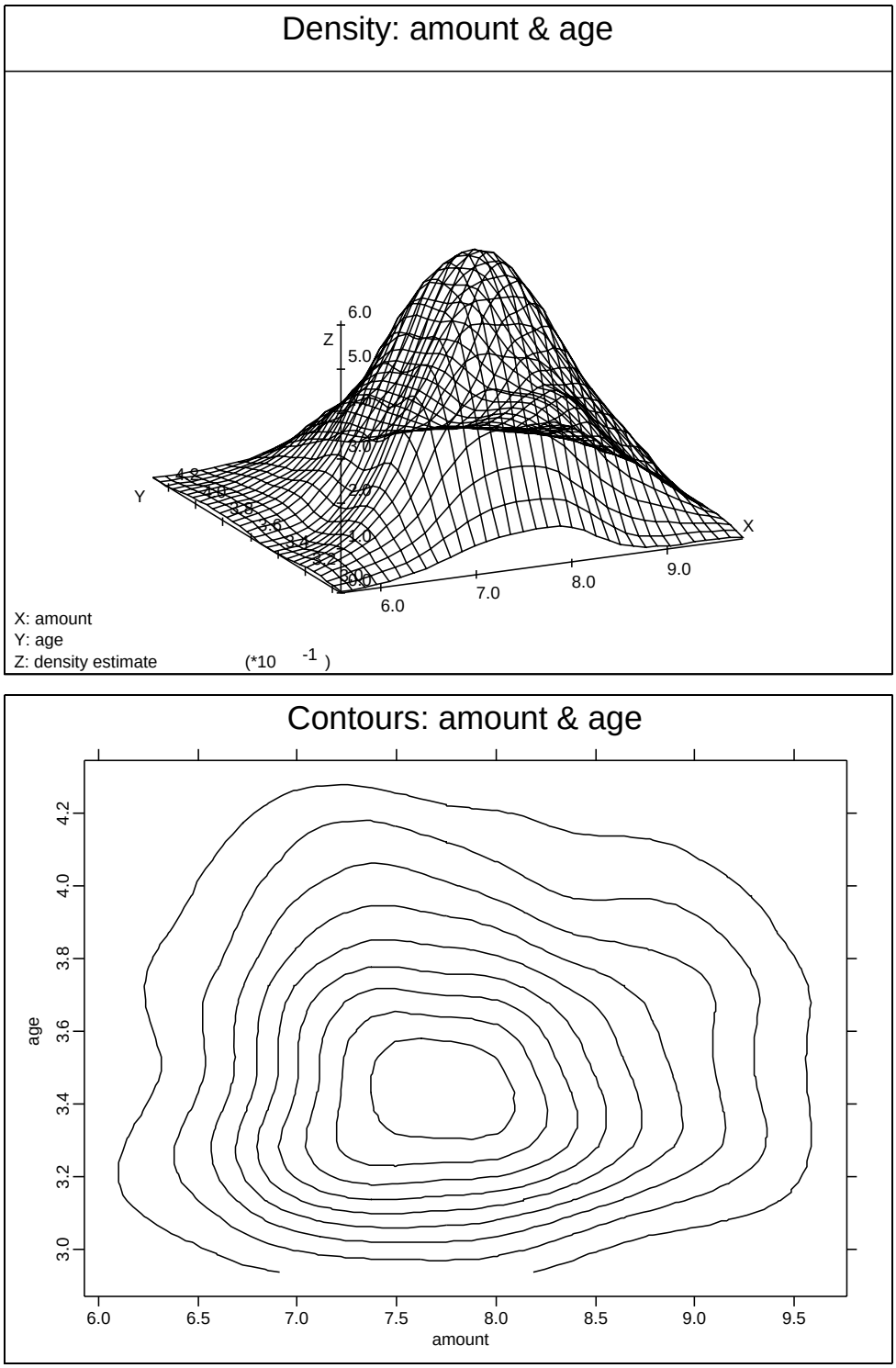


Figure 4: Two–dimensional density estimate (upper panel) and density contours (lower panel) for amount and age. $h_1 = 0.64$, $h_2 = 0.25$. Credit data, Fahrmeir and Hamerle (1984).

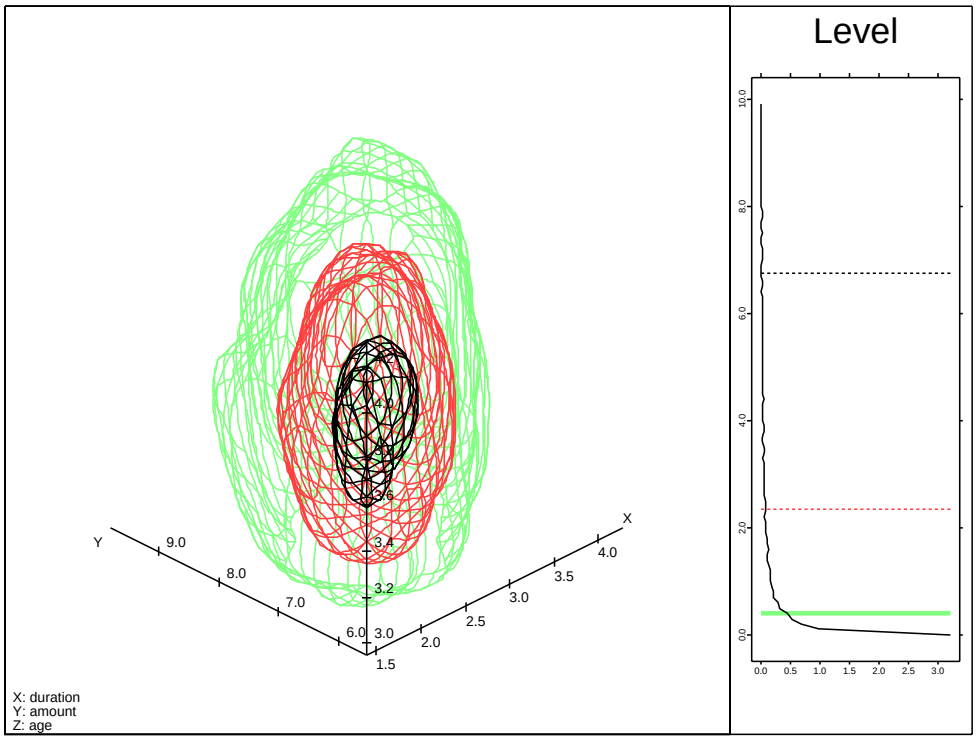


Figure 5: Three–dimensional density contours for duration, amount and age. $h_1 = 0.56$, $h_2 = 0.75$, $h_3 = 0.29$. Credit data, Fahrmeir and Hamerle (1984).

Quartic (Biweight) product kernel. Recall that the the univariate Quartic kernel is $K(u) = 0.9375(1-u^2)^2 \cdot \mathrm{I}(|u| \leq 1)$.

For three–dimensional density estimates, it is always possible to hold one variable fixed and to plot the density function only in dependence of the other variables. Alternatively, we can again plot contours of the density estimate, which now mean three-dimensional surfaces. Figure 5 shows this for the credit scoring variables. In the original version of this plot, red, green and blue surfaces show the values of the density estimate at the levels (in percent) indicated on the right. Colors and the possibility to rotate the contours on the computer screen eases the exploration of the data structures a lot. Of course, we are restricted to two-dimensional plots here. However, one can clearly recognize the ellipsoidal structure of the contour which indicates a relatively symmetric distribution.

1.2 Multivariate Kernel Regression

Multivariate nonparametric regression aims to estimate the functional relation between a response variable Y and a multivariate explanatory variable T, i.e. the conditional expectation

$$E(Y|T) = E\left(Y|T_1, \ldots, T_q\right) = m(T), \tag{20}$$

where as before $T = (T_1, \ldots, T_d)^T$. The relation

$$E(Y|T) = \int y f(y|t)\, dy = \frac{\int y f(y,t)\, dy}{f(x)}$$

leads by replacing the multivariate densities $f(y,t)$ by the kernel density estimate

$$\widehat{f}_{h,\mathbf{H}}(y,t) = \frac{1}{n}\sum_{i=1}^{n} K_h(Y_i - y)\, \mathcal{K}_{\mathbf{H}}(t_i - t)$$

and $f(t) = f_T(t)$ by (4) to the multivariate generalization of the Nadaraya–Watson estimator:

$$\widehat{m}_{\mathbf{H}}(t) = \frac{\sum\limits_{i=1}^{n} \mathcal{K}_{\mathbf{H}}\,(T_i - t)\, Y_i}{\sum\limits_{i=1}^{n} \mathcal{K}_{\mathbf{H}}\,(T_i - t)}. \tag{21}$$

Hence, the multivariate kernel regression estimator is just a weighted sum of the observed responses Y_i. The denominator ensures that the weights sum up to 1. Depending on the choice of the kernel, $\widehat{m}_{\mathbf{H}}(t)$ is a weighted average of those Y_i where T_i lies in a ball or cube around t.

Note that the multivariate Nadaraya–Watson estimator is a local constant estimator, i.e. the solution of

$$\min_{\beta_0} \sum_{i=1}^{n} \{Y_i - \beta_0\}^2\, \mathcal{K}_{\mathbf{H}}(T_i - t).$$

Replacing β_0 by a polynomial in $T_i - t$ yields a local polynomial kernel regression estimator. This definition of local polynomial kernel regression is a straightforward generalization of the univariate case. For details see Ruppert and Wand (1994). Let us illustrate this with the example of a local linear regression estimate. The minimization problem is here

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left\{Y_i - \beta_0 - (T_i - t)^T \beta_1\right\}^2 \mathcal{K}_{\mathbf{H}}(T_i - t).$$

The solution of the problem can hence equivalently be written as

$$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1^T)^T = \left(\mathbf{T}^T \mathbf{W} \mathbf{T}\right)^{-1} \mathbf{T}^T \mathbf{W} \mathbf{Y} \tag{22}$$

using the notations

$$\mathbf{T} = \begin{pmatrix} 1 & (T_1 - t)^T \\ \vdots & \vdots \\ 1 & (T_n - t)^T \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

and $\mathbf{W} = \operatorname{diag}\left(\mathcal{K}_{\mathbf{H}}(T_1 - t), \ldots, \mathcal{K}_{\mathbf{H}}(T_n - t)\right)$. In (22) $\widehat{\beta}_0$ estimates the regression function itself, whereas $\widehat{\beta}_1$ estimates the partial derivatives w.r.t. the components T. In the following we denote the multivariate local linear estimator as

$$\widehat{m}_{1,\mathbf{H}}(t) = \widehat{\beta}_0(t). \tag{23}$$

1.2.1 Bias, Variance and Asymptotics

The asymptotic conditional variance of the Nadaraya–Watson estimator $\widehat{m}_{\mathbf{H}}$ and the local linear $\widehat{m}_{1,\mathbf{H}}$ is identical and its derivation can be found in detail in Ruppert and Wand (1994):

$$Var\left\{\widehat{m}_{\mathbf{H}}(t)|T_1,\ldots,T_n\right\} = \frac{1}{n\det(\mathbf{H})}\left\|\mathcal{K}\right\|_2^2\frac{\sigma^2(t)}{f(t)}\left\{1+o_p(1)\right\}, \tag{24}$$

with $\sigma^2(t)$ denoting the variance function in $Var(Y|t)$.

We sketch the derivation of the asymptotic conditional bias since we find remarkable differences between both estimators. Denote M the second order Taylor expansion of $(m(T_1),\ldots,m(T_n))^T$, i.e.

$$M \approx m(t)\mathbb{1}_n + L(t) + \frac{1}{2}Q(t) = \mathbf{T}\begin{pmatrix} m(t) \\ \nabla_m(t)\end{pmatrix} + \frac{1}{2}Q(t), \tag{25}$$

with

$$L(t) = \begin{pmatrix} (T_1-t)^T\nabla_m(t) \\ \vdots \\ (T_n-t)^T\nabla_m(t)\end{pmatrix}, \quad Q(t) = \begin{pmatrix} (T_1-t)^T\mathcal{H}_m(t)(T_1-t) \\ \vdots \\ (T_n-t)^T\mathcal{H}_m(t)(T_n-t)\end{pmatrix}.$$

Additionally to (6) it holds

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)\,(T_i-t) &= \mu_2(\mathcal{K})\mathbf{H}\mathbf{H}^T\nabla_f(t) + o_p(\mathbf{H}\mathbf{H}^T\mathbb{1}_d),\\ \frac{1}{n}\sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)\,(T_i-t)(T_i-t)^T &= \mu_2(\mathcal{K})f(t)\mathbf{H}\mathbf{H}^T\nabla_f(t) + o_p(\mathbf{H}\mathbf{H}^T),\end{aligned}$$

see Ruppert and Wand (1994). Therefore the denominator of the conditional asymptotic expectation of the Nadaraya–Watson estimator $\widehat{m}_{\mathbf{H}}$ is approximately $f(t)$. Using $E(\mathbf{Y}|T_1,\ldots,T_n) = M$ and the Taylor expansion for M we have

$$\begin{aligned}&E\left\{\widehat{m}_{\mathbf{H}}|T_1,\ldots,T_n\right\}\\ &\approx \left\{f(t)+o_p(1)\right\}^{-1}\Big\{\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)m(t) + \sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)(T_i-t)^T\nabla_m(t)\\ &\qquad + \sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)(T_i-t)^T\mathcal{H}_m(t)(T_i-t)\Big\}\\ &\approx \left\{f(t)\right\}^{-1}\Big[f(t)m(t) + \mu_2(\mathcal{K})\nabla_m\mathbf{H}\mathbf{H}^T\nabla_f + \frac{1}{2}\mu_2(\mathcal{K})f(t)\,\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_m(t)\mathbf{H}\}\Big].\end{aligned}$$

This is summarized in the following theorem.

THEOREM 1

The conditional asymptotic bias and variance of the multivariate Nadaraya–Watson kernel

regression estimator are

$$E\left\{\widehat{m}_{\mathbf{H}}|T_1,\ldots,T_n\right\} - m(t) \approx \mu_2(\mathcal{K})\frac{\nabla_m(t)^T\mathbf{H}\mathbf{H}^T\nabla_f(t)}{f(t)} + \frac{1}{2}\mu_2(\mathcal{K})\, tr\{\mathbf{H}^T\mathcal{H}_m(t)\mathbf{H}\}$$

$$Var\left\{\widehat{m}_{\mathbf{H}}|T_1,\ldots,T_n\right\} \approx \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2\frac{\sigma^2(t)}{f(t)}$$

in the interior of the support of f_T.

Recall the notation $e_1 = (1,0,\ldots,0)^T$ for the first unit vector in $I\!R^d$. Then we can write the local linear estimator as

$$\widehat{m}_{1,\mathbf{H}}(t) = e_1^T\left(\mathbf{T}^T\mathbf{W}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{W}\mathbf{Y}.$$

Now we have using (22) and (25)

$$\begin{aligned} & E\left\{\widehat{m}_{1,\mathbf{H}}|T_1,\ldots,T_n\right\} - m(t) \\ &= e_1^T\left(\mathbf{T}^T\mathbf{W}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{W}\mathbf{T}\left\{\begin{pmatrix} m(t) \\ \nabla_m(t)\end{pmatrix} + \frac{1}{2}Q(t)\right\} - m(t) \\ &= \frac{1}{2}e_1^T\left(\mathbf{T}^T\mathbf{W}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{W}Q(t) \end{aligned}$$

since $e_1^T[m(t),\nabla_m(t)^T] = m(t)$. Hence, the numerator of the asymptotic conditional bias only depends on the quadratic term. This is one of the key points in asymptotics for local polynomial estimators. If we would use local polynomials of order d and expand M up to order $d+1$, then only the term of order $d+1$ would appear in the numerator of the asymptotic conditional bias. Of course this to be paid with a more complicated structure of the denominator.

THEOREM 2

The conditional asymptotic bias and variance of the multivariate local linear regression estimator are

$$E\left\{\widehat{m}_{1,\mathbf{H}}|T_1,\ldots,T_n\right\} - m(t) \approx \frac{1}{2}\mu_2(\mathcal{K})\, tr\{\mathbf{H}^T\mathcal{H}_m(t)\mathbf{H}\}$$

$$Var\left\{\widehat{m}_{1,\mathbf{H}}|T_1,\ldots,T_n\right\} \approx \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2\frac{\sigma^2(t)}{f(t)}$$

in the interior of the support of f_T.

For all omitted details on the proof of Theorem 2 we refer again to Ruppert and Wand (1994). They also point out that the local linear estimate has same order conditional bias in the interior as well as in the boundary of the support of f_T. Fan, Gasser, Gijbels, Brockmann and Engel (1993) point out that the multivariate local linear fit with Epanechnikov kernel is a best linear estimator and has a minimax efficiency of at least 89.4% among all estimators.

1.2.2 Bandwidth Selection and Practical Aspects

Principally, the methods to choose a smoothing parameter in nonparametric regression are the same as in density estimation. Again, plug–in and resampling ideas are employed for finding a global bandwidth $\mathbf{H}$ or a local bandwidth $\mathbf{H}(t)$.

For our presentation, we concentrate on the classical cross–validation bandwidth selector. As a motivation, we introduce the *residual sum of squares* (RSS) as a (naive) way to asses the goodness of fit

$$RSS(\mathbf{H}) = n^{-1} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{\mathbf{H}}(X_i)\}^2 , \tag{26}$$

which is also called resubstitution estimate for the *averaged squared error* (ASE). Note, that we concentrate on the Nadaraya–Watson estimator in the moment.

There is a problem with the RSS: Y_i is used in $\widehat{m}_{\mathbf{H}}(X_i)$ to predict itself. As a consequence, $ASE(\mathbf{H})$ can be made arbitrarily small by letting $\mathbf{H} \to 0$ (in which case $\widehat{m}_{\mathbf{H}}$ is an interpolation of the Y_i's). This leads to the *cross–validation* function

$$CV(\mathbf{H}) = n^{-1} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)\}^2 . \tag{27}$$

This function replaces $\widehat{m}_{\mathbf{H}}(X_i)$ in (26) with the *leave–one–out*-estimator

$$\widehat{m}_{\mathbf{H},-i}(X_i) = \frac{\sum_{j \neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j) Y_j}{\sum_{j \neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j)}. \tag{28}$$

and is equivalent to a different approach, which multiplies each term in $RSS(\mathbf{H})$ by a *penalizing function* that is correcting for the downward bias of the resubstitution estimate. For the Nadaraya–Watson estimator

$$\begin{aligned} CV(\mathbf{H}) &= \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)\}^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{\mathbf{H}}(X_i)\}^2 \left\{ \frac{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)}{Y_i - \widehat{m}_{\mathbf{H}}(X_i)} \right\}^2 \end{aligned} \tag{29}$$

and

$$\begin{aligned} \frac{Y_i - \widehat{m}_{\mathbf{H}}(X_i)}{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)} &= \frac{\sum\limits_j \mathcal{K}_{\mathbf{H}}(X_i - X_j) Y_j - Y_i \sum\limits_j \mathcal{K}_{\mathbf{H}}(X_i - X_j)}{\sum\limits_{j \neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j) Y_j - Y_i \sum\limits_{j \neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j)} \cdot \frac{\sum\limits_{j \neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j)}{\sum\limits_j \mathcal{K}_{\mathbf{H}}(X_i - X_j)} \\ &= 1 - \frac{\mathcal{K}_{\mathbf{H}}(0)}{\sum\limits_j \mathcal{K}_{\mathbf{H}}(X_i - X_j)}. \end{aligned} \tag{30}$$

Therefore the cross–validation approach is equivalent to the penalizing functions concept and shares the same asymptotic properties. Note that (30) is a function of the i–th

diagonal element of the smoother matrix. More precisely, cross–validation is equivalent with *generalized cross–validation* (Craven and Wahba, 1979) in this case. Härdle, Hall and Marron (1988) show asymptotic optimality of the selected bandwidth, the rate of convergence is slow though. An improved bandwidth selection is discussed in Härdle, Hall and Marron (1992).

We want to remark that (29) and (30) also imply that the computation of $CV(\mathbf{H})$ needs actually not more computational effort than the computation of $m_{\mathbf{H}}(X_1), \ldots, m_{\mathbf{H}}(X_n)$. However, the optimization over a matrix $\mathbf{H}$ might be cumbersome, hence diagonal bandwidth matrices (or even $\mathbf{H} = h\mathbf{I}_q$ with appropriate standardization of the data) are still preferred in practice.

Before we consider cross–validation bandwidth selection in the local linear case, we want to comment on the practical computation of the estimator. Principally, since multivariate kernel regression estimators can be expressed as local polynomial estimators, their computation can be done by any statistical package that is able to run weighted least squares regression. However, since we estimate a function, this weighted least squares regression has to be performed in all observation points or on a grid of points in $I\!R^q$. Therefore, explicit formulae are useful.

We will give an formula for the multivariate local linear estimator in the following. Consider for a fixed point t the sums

$$\begin{aligned}
\mathcal{S}_0 = \mathcal{S}_0(t) &= \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(T_i - t) \\
\mathcal{S}_1 = \mathcal{S}_1(t) &= \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(T_i - t)(T_i - t) \\
\mathcal{S}_2 = \mathcal{S}_2(t) &= \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(T_i - t)(T_i - t)(T_i - t)^T \\
\mathcal{T}_0 = \mathcal{T}_0(t) &= \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(T_i - t)Y_i \\
\mathcal{T}_1 = \mathcal{T}_1(t) &= \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(T_i - t)(T_i - t)Y_i.
\end{aligned}$$

Note that $\mathcal{S}_1$ and $\mathcal{T}_1$ are q–variate vectors and that $\mathcal{S}_2$ is a $q \times q$ matrix. Then for the local linear estimate we can write

$$\widehat{\beta} = \begin{pmatrix} \mathcal{S}_0 & \mathcal{S}_1^T \\ \mathcal{S}_1 & \mathcal{S}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{T}_0 \\ \mathcal{T}_1 \end{pmatrix}. \tag{31}$$

For the regression function we need only the first component $e_1^T \widehat{\beta}$. Applying block–wise matrix inversion we obtain

$$e_1^T \begin{pmatrix} \mathcal{S}_0 & \mathcal{S}_1^T \\ \mathcal{S}_1 & \mathcal{S}_2 \end{pmatrix}^{-1} = \left(\mathcal{S}_0 - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{S}_1\right)^{-1} \cdot \begin{pmatrix} 1 & -\mathcal{S}_1^T \mathcal{S}_2^{-1} \end{pmatrix}$$

and hence

$$\widehat{m}_{1,\mathbf{H}}(t) = \frac{\mathcal{T}_0 - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{T}_1}{\mathcal{S}_0 - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{S}_1}. \tag{32}$$

The cross–validation criterion here is a weighted RSS as in (29). If we denote the leave–one–out estimator $\widehat{m}_{1,\mathbf{H},-i}(t)$ and define its components accordingly, we observe

$$\begin{aligned} \mathcal{S}_{0,-i} &= \mathcal{S}_0 - \mathcal{K}_{\mathbf{H}}(0), \quad \mathcal{S}_{1,-i} = \mathcal{S}_1, \quad \mathcal{S}_{2,-i} = \mathcal{S}_2 \\ \mathcal{T}_{0,-i} &= \mathcal{T}_0 - Y_i \mathcal{K}_{\mathbf{H}}(0), \quad \mathcal{T}_{1,-i} = \mathcal{T}_1. \end{aligned}$$

This means

$$\widehat{m}_{1,\mathbf{H},-i}(t) = \frac{\mathcal{T}_0 - Y_i \mathcal{K}_{\mathbf{H}}(0) - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{T}_1}{\mathcal{S}_0 - \mathcal{K}_{\mathbf{H}}(0) - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{S}_1}$$

which yields in analogy to (30)

$$\frac{Y_i - \widehat{m}_{\mathbf{H}}(X_i)}{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)} = 1 - \frac{\mathcal{K}_{\mathbf{H}}(0)}{\mathcal{S}_0 - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{S}_1}. \tag{33}$$

As in the Nadaraya–Watson case, (33) is a function of the i–th diagonal element of the smoother matrix. A summary of bandwidth selection methods other than cross–validation can be found in particular in Fan and Gijbels (1995). They also cover rule–of–thumb approaches.

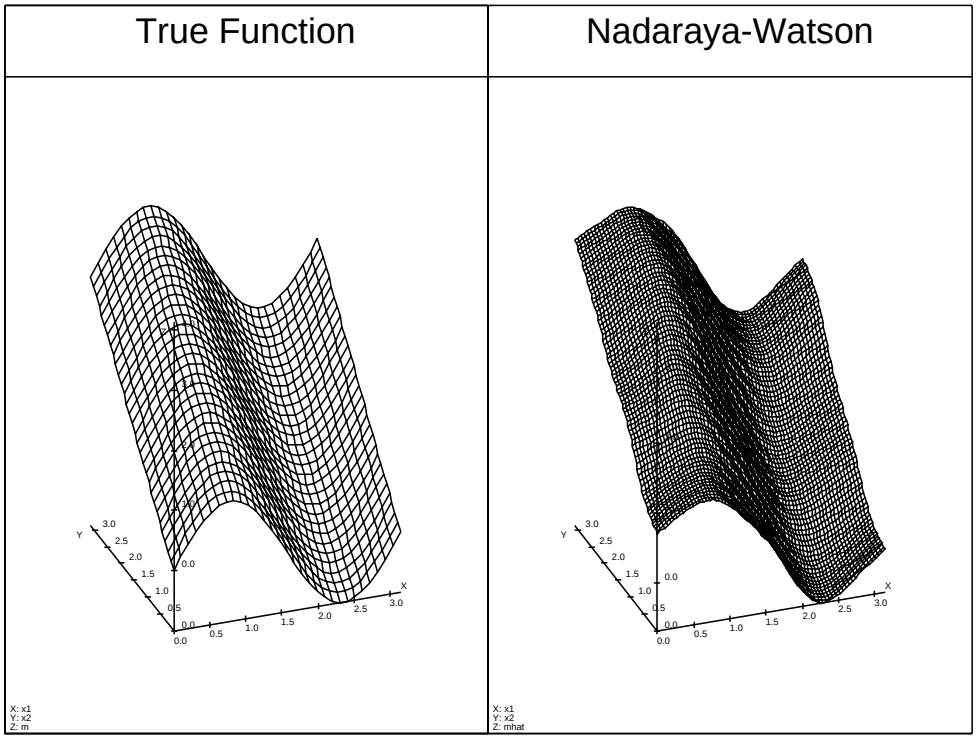


Figure 6: Two–dimensional Nadaraya–Watson Estimate.

Recall that (32) estimates the regression function only in one point t. To estimate the regression plane we have to apply (32) on a two–dimensional grid of points. The WARPing technique (binning) described in Härdle and Scott (1992) and applied to local polynomial kernel regression by Fan and Marron (1994), Fan and Müller (1995), can be used to speed up calculations. See also Wand (1994) for an analysis of fast computation methods for multivariate kernel estimation.

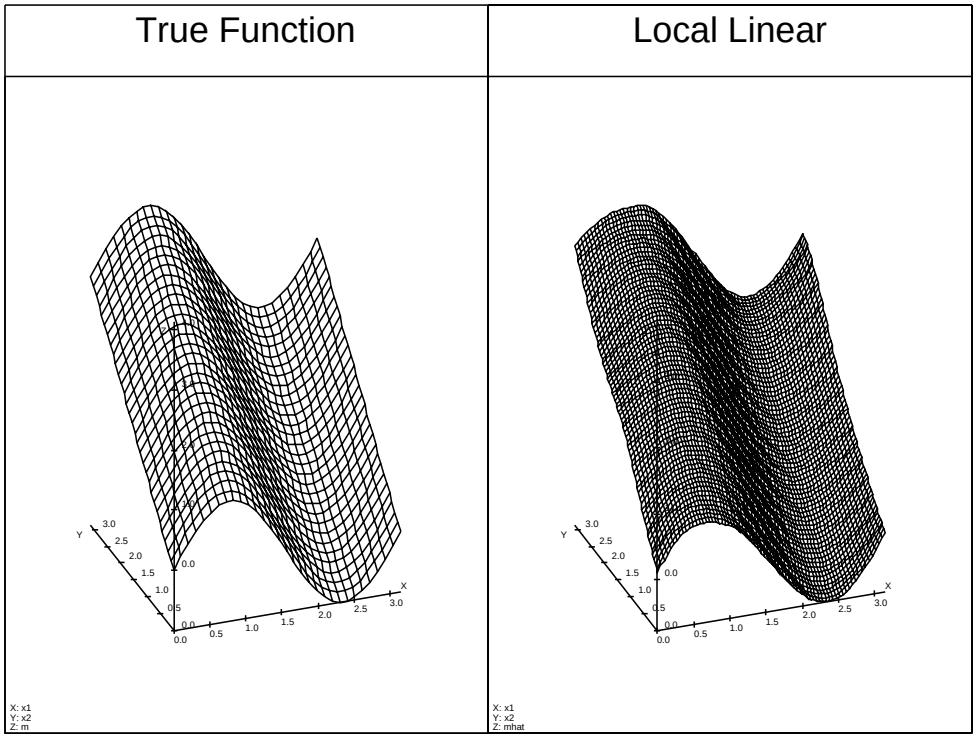


Figure 7: Two–dimensional Local Linear Estimate.

Figures 6 and 7 show the bivariate Nadaraya–Watson and local linear estimate for simulated data. The underlying curve is in fact an additive combination of a sine function in the first and a linear function in the second argument. Note, that we have chosen the same bandwidth in both estimates.

Of course, nonparametric kernel regression estimation is not limited to bivariate distributions. A practical issue is the graphical display for higher dimensional multivariate functions. This was already considered when we discussed the graphical representation of multivariate density estimates. The corresponding remarks apply here again. The general problem in multivariate nonparametric estimation is the *curse of dimensionality*. Recall that the nonparametric regression estimators are based on the idea of local (weighted) averaging. In higher dimensions the observations are usually sparsely distributed for reasonable sample sizes, and consequently estimators based on local averaging perform unsatisfactorily in this situation.

Technically, one can explain this effect by looking at the *AMISE* again. Consider a multivariate regression estimator with the same bandwidth h for all components, e.g. a Nadaraya–Watson or local linear estimator with bandwidth matrix $\mathbf{H} = h\mathbf{I}_q$. Here the asymptotic *MISE* also depends on q:

$$AMISE(n,h) = \frac{1}{nh^q}C_1 + h^4 C_2.$$

where C_1 and C_2 are constants that neither depend on n nor h. If we derive the optimal bandwidth we find that $h_{opt} \sim n^{-1/(4+q)}$ and hence the rate of convergence for *AMISE* is $n^{-4/(4+q)}$. One can clearly see that the speed of convergence decreases dramatically for higher dimensions q.

2 Semiparametric Generalized Regression Models

As the name suggests, semiparametric models combine two elements, one of them to be estimated nonparametrically, the other one requiring the estimation of a set of finite dimensional parameters. In this section we concentrate on single index and generalized partial linear models.

Often a canonical partitioning of the explanatory variables exists. In particular, if there are binary or discrete explanatory variables we keep them separate from the other design variables. In the following we denote by $T = (T_1, \ldots, T_q)^T$ a vector of continuous explanatory variables and refer to $X = (X_1, \ldots, X_p)^T$ as the discrete part of the variables.

Semiparametric generalized linear models are widely used in modeling binary choice, i.e. in situations where the response variable has two alternatives. Recall the example on credit scoring which was introduced previously. In the analysis of discrete response variables one typically models the expected value of the response as a nonlinear monotone function of a linear combination of the explanatory variables. Examples are probit or logit models where the nonlinear (link) function is the cumulative distribution function of a normal respectively logistic distribution, see McCullagh and Nelder (1989). Then the so-called *generalized linear model* has the form

$$E(Y|X,T) = G(X^T\beta + T^T\gamma), \tag{34}$$

with a known monotone function G and an unknown parameters β and γ. The model (34) combines computational feasibility (especially for discrete covariates) with good interpretability of the "index" $X^T\beta + T^T\gamma$ and therefore has found wide application in all fields of applied statistics, see e.g. Fahrmeir and Tutz (1994), Maddala (1983). However, for some applications it may be argued that the assumption of (34) is too restrictive

(Horowitz, 1993). Indeed it may be not even clear if the relationship between the influential variables and the response is monotone.

Several approaches have been proposed to generalize parametric regression models in order to allow nonmonotone relationships between explanatory variables and the dependent variable Y. We will focus on two classes of semiparametric models that have received a lot of attention.

- Generalization of the known (parametric) link function G to an unknown (nonparametric) link function $g(\bullet)$ yields the *single index model* (SIM)

$$E(Y|X,T) = g(X^T\beta + T^T\gamma),$$

  also called a *one term projection pursuit model* in statistics. Obviously, due to the nonparametric character of the link function conventional parametric estimation procedures can no longer be applied in this case. Instead, nonparametric estimators will now be necessary. In this chapter we give an overview how this model can be estimated using kernel methods.

- Generalization of the linear form $X^T\beta + T^T\gamma$ to a partial linear form $X^T\beta + m(T)$ yields the *generalized partial linear model* (GPLM)

$$E(Y|X,T) = G\left\{X^T\beta + m(T)\right\},$$

  G denoting a known link function as in the GLM model. Here, the $m(\bullet)$ will be a multivariate nonparametric function of the variable T.

  In high dimensions of T the estimate of the nonparametric function $m(\bullet)$ faces the same problems as the fully nonparametric multidimensional regression function estimates: the curse of dimensionality and the practical problem of interpretability.

  Hence it might be reasonable to think about a lower dimensional nonparametric modelization of the nonparametric part. A possible alternative is the GPLM with an additive structure in the nonparametric component, i.e. the *generalized additive model* (GAM).

$$E(Y|X,T) = G\left\{X^T\beta + m_1(T_1) + \ldots + m_d(T_d)\right\}.$$

  Here, the $m_j(\bullet)$ will be univariate nonparametric functions of the variables T_j.

Formally, we can summarize these generalizations as shown in Table 2. The last entry in this table is empty because we do not know (yet) of any literature which deals exactly with this situation. Of course, there is a number of approaches which attempt to fill this gap: as e.g. neural networks, sliced inverse and projection pursuit regression.

components \ link	known	unknown
linear	GLM	SIM
partial nonparametric	GPLM	

Table 2: Parametric → Semiparametric

2.1 Generalizing the link function: Single Index Models

Single index models derive their name from the economic term "index", a summary of different variables into one number. Hence, if it is possible to summarize all information in one single number this is to be called a single index. Meanwhile, there has been a number of methods proposed do deal with these models. A straightforward semiparametric GLM extension is provided by Weisberg and Welsh (1994). They estimated the unknown link function and its derivative (for the Fisher scoring algorithm) by an kernel smoother. Ichimura (1993) uses a similar idea within a least squares criterion. Klein and Spady (1993) show an asymptotic efficiency result for a pseudo–likelihood binary choice estimator.

All these three methods require optimization of a pseudo–likelihood of possibly complicated structure. We present here an direct approach which avoids numerical iterations. The estimation of the single index model

$$E(Y|X,T) = g(X^T\beta + T^T\gamma) \tag{35}$$

is carried out in two steps. First the coefficients vectors β, γ are estimated, then using the obtained index values $X_i^T\widehat{\beta} + T_i^T\widehat{\gamma}$ one can estimate g by usual univariate nonparametric regression.

2.1.1 Average Derivative Estimation

Consider for a moment only the continuous part of the variables, $T = (T_1, \ldots, T_q)^T$. Denote the regression function to estimate by $m(\bullet)$, i.e. $E(Y|T) = m(T)$. The vector of average derivatives is given by

$$\delta = E\left\{\nabla_m(T)\right\} = E\left\{g'(T^T\beta)\right\}\ \beta, \tag{36}$$

where $\nabla_m(t)$ is the vector of partial derivatives of $m(\bullet)$ and g' the derivative of $g(\bullet)$.

Looking at (36) shows that δ equals β up to scale. Hence, any estimate of δ determines β up to scale. The estimation of δ can be carried out by means of several *average derivative estimation* (ADE) methods. We will concentrate on estimators based on the density

function of T, however a variety of other methods exist. For an overview see Stoker (1991).

The key idea on ADE based on the density $f(\bullet)$ of T lies in "transferring" the derivative of the regression function m on to the derivative of the density f. Consider

$$\delta = E\{\nabla_m(T)\} = \int \nabla_m(t) f(t)\, dt. \tag{37}$$

Partial integration yields $E\{\nabla_m(T)\} = -\int m(t)\nabla_f(t)\, dt$ if $f(t)\, m(t) \to 0$ is assumed for $\|t\| \to \infty$. Hence by introducing the *score vector*

$$\ell(t) = \nabla_{\log f}(t) = \frac{\nabla_f(t)}{f(t)} \tag{38}$$

one arrives at

$$\delta = -\int \frac{\nabla_f(t)}{f(t)}\, f(t)\, m(t)\, dt = \int \ell(t)\, m(t)\, f(t)\, dt = E\{\ell(T)\, m(T)\}. \tag{39}$$

Employing $E\{\ell(T)\, m(T)\} = E\{\ell(T)\, Y\}$ immediately allows to estimate δ by the sample analog

$$\widehat{\delta} = n^{-1} \sum_{i=1}^{n} \widehat{\ell}_{\mathbf{H}}(T_i)\, Y_i, \tag{40}$$

where $\ell(t)$ is approximated by $\widehat{\ell}_{\mathbf{H}}(t) = -\{\widehat{f}_{\mathbf{H}}(t)\}^{-1} \left(-\partial_1 \widehat{f}_{\mathbf{H}}(t), \ldots, -\partial_q \widehat{f}_{\mathbf{H}}(t)\right)$. Here, $\widehat{f}_{\mathbf{H}}(t)$ is the multivariate kernel density estimator and $\partial_j \widehat{f}_{\mathbf{H}}(t)$ are the partial derivatives of this multivariate kernel density estimator (which are used for estimating the partial derivatives of the density)

$$\partial_j \widehat{f}_{\mathbf{H}}(t) = \frac{1}{n \det(\mathbf{H})} \sum_{j=1}^{n} \partial_j K_{\mathbf{H}}\,(t - T_j) \tag{41}$$

Due to the sparseness of data in high dimensions, the use of $\widehat{f}_{\mathbf{H}}$ can also be problematic since $\widehat{\ell}_{\mathbf{H}}$ might behave bad in regions of small density. Hence Härdle and Stoker (1989) propose to use the ADE estimator

$$\widehat{\delta} = n^{-1} \sum_{i=1}^{n} \widehat{\ell}_{\mathbf{H}}(T_i)\, Y_i\ \mathrm{I}\{\widehat{f}_{\mathbf{H}}(T_i) > b_n\}, \tag{42}$$

where $\mathrm{I}\{\widehat{f}_{\mathbf{H}}(T_i) > b_n\}$ is an indicator that excludes too small density values. The trimming bounds b_n are chosen such that $b_n \to 0$ for $n \to \infty$.

Regarding the sampling distribution of the estimator, $\widehat{\delta}$ Härdle and Stoker (1989) have shown that

$$\sqrt{n}(\widehat{\delta} - \delta) \xrightarrow[n\to\infty]{\mathcal{L}} N(\mathbf{0}, \mathbf{\Sigma}_\delta),$$

where $\mathbf{\Sigma}_\delta$ is the covariance matrix of $\ell(T)Y + \{\nabla_m(T) - \ell(T)m(T)\}$. Note that $\widehat{\delta}$ achieves $\sqrt{n}$-convergence, a rate that is typically achieved by parametric estimators.

The need for "trimming" the ADE is one of the problems associated with a random denominator. Random denominators also complicate the derivation of the distributional properties. These difficulties are overcome by *density weighted average derivative estimation* (WADE) of Powell, Stock and Stoker (1989). Observe that the density weighted average derivative shares the property of the (unweighted) average derivative of being proportional to the coefficient vector β in index models:

$$\delta = E\left\{\nabla_m(T)\, w(T)\right\} = E\left\{g'(T^T\beta)\, w(T)\right\}\ \beta, \tag{43}$$

A "natural" weight function is given by the density f itself. Calculations similar to those for the unweighted ADE give with $w(t) = f(t)$

$$\begin{aligned}\delta &= \int \nabla_m(t)\, f^2(t)\, dt = -2\int m(t)\nabla_f(t)\, f(t)\, dt \\ &= -2\, E\{Y\ \nabla_f(T)\}.\end{aligned}$$

Thus one may estimate β up to scale by

$$\widehat{\delta} = -\frac{2}{n}\sum_{i=1}^{n} Y_i\, (\partial_1 \widehat{f}_{\mathbf{H}}(t), \ldots, \partial_q \widehat{f}_{\mathbf{H}}(t))^T. \tag{44}$$

The WADE estimator defined in (44) shares the desirable distributional features of the ADE estimator ($\sqrt{n}$-consistency, asymptotic normality) while not requiring any trimming in practice.

Finally, an estimate for $g(\bullet)$ can be found by applying an univariate estimation method to $\widehat{\delta}^T T_i$ and Y_i. Härdle and Stoker (1989) showed for the Nadaraya–Watson estimator the usual rate of convergence $\sqrt{nh}$, when $h \sim n^{-1/5}$.

2.1.2 Including Discrete Explanatory Variables

By definition, derivatives can only be calculated if the variable under study is continuous. Thus, the method of weighted or unweighted ADE fails when discrete variables $X = (X_1, \ldots, X_d)^T$ needs to be included into the model. Before giving a more general solution, let us explain how the coefficient of one dichotomous variable presents in the model. Recall the SIM

$$E(Y|T, X) = g(X^T\beta + T^T\gamma)$$

with T the continuous and X the discrete part of the covariates. In the simplest case, we suppose that is X is binary, i.e. either $X = 1$ or $X = 0$. Then, this model can be "split" into two submodels

$$\begin{aligned}E(Y|T, X) &= g(T^T\gamma) \qquad \text{if } X = 0 \\ E(Y|T, X) &= g(T^T\gamma + \beta) \quad \text{if } X = 1.\end{aligned}$$

These are in fact two models to be estimated, one for $X = 0$ and one for $X = 1$. Note that γ alone could be estimated from the first equation only.

Theoretically, the same T_i can be associated with either $X_i = 0$ yielding an index value of $\gamma^T T_i$ or with $X_i = 1$ leading to an index value of $\gamma^T T_i + \beta$. Thus the difference between the two indices is exactly β. In practice finding these *horizontal* differences will be rather difficult. A common approach parts from the observation that the *integral* difference between the two link functions also equals β. A very simple estimator is proposed in Korostelev and Müller (1995). Essentially, the coefficient of the binary explanatory variable can be estimated by

$$\widehat{\beta} = \widehat{J}^{(1)} - \widehat{J}^{(0)}$$

with

$$\widehat{J}^{(0)} = \sum_{i=0}^{n_0} \gamma^T (T_{i+1}^{(0)} - T_i^{(0)}) Y_i^{(0)}, \quad \widehat{J}^{(1)} = \sum_{i=0}^{n_1} \gamma^T (T_{i+1}^{(1)} - T_i^{(1)}) Y_i^{(1)},$$

where the superscripts $^{(0)}$ and $^{(1)}$ denote the observations coming from the subsamples according to $X_i = 0$ and $X_i = 1$. The estimator is in the simplest case of a binary Y variable $\sqrt{n}$-consistent and can be improved for efficiency by a one–step estimator, see Korostelev and Müller (1995).

Horowitz and Härdle (1996) extend this approach to multivariate multi–categorical X and arbitrary range of Y. Again, this approach is based on a split of the whole sample into subsamples according to the categories of X. Consider the thresholded link function

$$\tilde{g} = c_o \, \mathrm{I}(g < c_o) + g \, \mathrm{I}(c_o \le g \le c_1) + c_1 \, \mathrm{I}(g > c_1).$$

Denote $x^{(k)}$ a possible realization of X, then the integrated link function conditional on $x^{(k)}$ is

$$J^{(k)} = \int_{v_o}^{v_1} \tilde{g}(v + \beta^T x^{(k)}) \, dv.$$

Now compare the integrated link functions for all X–categories $x^{(k)}$ $(k = 1, \ldots, M)$ to the first X–category $x^{(0)}$. It holds

$$J^{(k)} - J^{(0)} = (c_1 - c_0) \left\{ x^{(k)} - x^{(0)} \right\} \beta,$$

hence with

$$\Delta J = \begin{pmatrix} J^{(1)} - J^{(0)} \\ \cdots \\ J^{(M)} - J^{(0)} \end{pmatrix}, \; \Delta x = \begin{pmatrix} x^{(1)} - x^{(0)} \\ \cdots \\ x^{(M)} - x^{(0)} \end{pmatrix}$$

one gets $\Delta J = (c_1 - c_0) \, \Delta x \, \beta$. This yields finally

$$\beta = (c_1 - c_o)^{-1} (\Delta x^T \Delta x)^{-1} \Delta x^T \Delta J \tag{45}$$

to determine β. The estimation of β is based on replacing $J^{(k)}$ in 45 by

$$\widehat{J}^{(k)} = \int_{v_o}^{v_1} \widehat{\widetilde{g}}(v + \beta^T x^{(k)})\, dv$$

with $\widehat{\widetilde{g}}$ a nonparametric estimate of the thresholded link function $\widetilde{g}$. This estimator is obtained by a univariate regression of the estimated "continuous" indices $\widehat{\gamma}^T T_i^{(k)}$ on $Y_i^{(k)}$. Horowitz and Härdle (1996) show that using a $\sqrt{n}$–consistent estimate $\widehat{\gamma}$ and a Nadaraya–Watson estimator $\widehat{\widetilde{g}}$ the estimated coefficient $\widehat{\beta}$ is itself $\sqrt{n}$–consistent and has an asymptotic normal distribution.

2.2 Generalizing the index: Generalized Partial Linear Models

An alternative way to incorporate an nonmonotone dependence of the response on the continuous variables is given by a *generalized partial linear model* (GPLM)

$$E(Y|X,T) = G\{X^T\beta + m(T)\}, \tag{46}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a finite dimensional parameter and $m(\bullet)$ is a smooth function. These models allow a nonparametric inclusion of a part of the explanatory variables. In practice this might be only those continuous variables which have most influence on the dependent variable Y. In this section we will deal with the GPLM in general and shortly with *generalized partial linear partial additive models* (GAM).

Estimators for β and $m(\bullet)$ have been proposed by Severini and Wong (1992), Severini and Staniswalis (1994) and Hunsberger (1994). Carroll, Fan, Gijbels and Wand (1995) proposed an extension to *generalized partial linear single index model* (GPLSIM) which uses a single index model instead of the fully nonparametric function $m(\bullet)$.

2.2.1 Semiparametric Maximum Likelihood

We sketch the approach of Severini and Wong (1992) and Severini and Staniswalis (1994) which use two different likelihood functions for the estimation of the parametric and semiparametric components. The estimation of model (46) is computationally feasible by the idea that an estimate $\widehat{\beta}$ can be found for known m, and an estimate $\widehat{m}$ can be found for known β. Define

$$\begin{aligned} \mu &= E(Y|X,T) = G\{\beta^T X + m(T)\} \\ \sigma^2 V(\mu) &= Var(Y|X,T) \end{aligned}$$

and denote by $\ell(\mu, y)$ the individual log–likelihood or quasi–likelihood function (if the distribution of Y does not belong to an exponential family).

The "parametric" likelihood function

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \ell \left[G\{X_i^T \beta + m_\beta(T_i), Y_i\} \right] \tag{47}$$

is used to obtain $\widehat{\beta}$. A "smoothed" or "local" likelihood

$$\mathcal{L}^S(\eta) = \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(t - T_i)\, \ell \left\{ G(X_i^T \beta + \eta, Y_i) \right\} \tag{48}$$

is optimized to estimate the smooth function $m_\beta(t) = \eta$ at point t. Note that the use of this smoothed likelihood function leads to the equivalent of the Nadaraya–Watson estimator $\widehat{m}_{\mathbf{H}}$ in ordinary regression. To obtain a local polynomial estimator of the nonparametric part $m(\bullet)$ we need to incorporate polynomial terms into the smoothed likelihood. In the local linear case we would use

$$\mathcal{L}^S(\eta_0, \eta_1) = \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(t - T_i)\, \ell \left[G\{X_i^T \beta + \eta_0 + (T_i - t)^T \eta_1, Y_i\} \right] \tag{49}$$

and get $m_\beta(t) = \eta_0$ at point t. Analogous to local linear regression η_1 points to the gradient of $m(\bullet)$ in t.

The computational algorithm consists in searching maxima of both likelihoods simultaneously. We stay in the framework of an Nadaraya–Watson type estimation of m. Severini and Staniswalis (1994) show that the resulting estimator $\widehat{\beta}$ is $\sqrt{n}$–consistent and asymptotically normal, and that estimators $\widehat{m} = \widehat{m}_{\widehat{\beta}}$ are consistent in supremum norm. Note that m is estimated as a function of the parametric component β which yields an asymptotically efficient estimate $\widehat{\beta}$ (Severini and Wong, 1992). The possible scale parameter σ can be estimated by

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{\mu}_i)^2 / V(\widehat{\mu}_i), \tag{50}$$

where $\widehat{\mu}_i = G\{\widehat{\beta}^T X_i + \widehat{m}(T_i)\}$.

The algorithm which we will present here corresponds to that proposed in Severini and Staniswalis (1994) for some special cases of link function and distributions of Y. In order to avoid boundary effects, one can use a weight function in the convergence criterion or trimming in the estimation of β as in Severini and Staniswalis (1994).

Define $\eta_j(\beta) = \widehat{m}_\beta(t_j)$ and $\ell_i(u) = \ell\{(G(u), Y_i\}$. For example, in a binary response model we have $\ell_i(u) = Y_i \log G(u) + (1 - Y_i) \log\{1 - G(u)\}$. In the following, ℓ_i' and ℓ_i'' denote the derivatives of $\ell_i(u)$ with respect to u. The maximization of the smoothed quasi–likelihood (48) requires to solve

$$0 = \sum_{i=1}^{n} \ell_i'\{X_i^T \beta + \eta_j(\beta)\} \mathcal{K}_{\mathbf{H}}(T_i - T_j). \tag{51}$$

In some models (in particular for identity and exponential link functions G) equation (51) can be solved explicitly for $\eta_j(\beta)$. Differentiation of (51) leads to an estimate for η_j' as a function of β

$$\eta_j'(\beta) = \frac{-\sum_{i=1}^{n} \ell_i''\{X_i^T\beta + \eta_j(\beta)\}\mathcal{K}_{\mathbf{H}}(T_i - T_j)X_i}{\sum_{i=1}^{n} \ell_i''\{X_i^T\beta + \eta_j(\beta)\}\mathcal{K}_{\mathbf{H}}(T_i - T_j)}. \tag{52}$$

For β we have to solve

$$0 = \sum_{i=1}^{n} \ell_i'\{X_i^T\beta + \eta_i(\beta)\}\,\{X_i + \eta_i'(\beta)\}. \tag{53}$$

Equations (51)–(53) imply the following iterative Newton–Raphson type algorithm to find $\widehat{\beta}$ and $\widehat{m}(t_j) = \widehat{\eta}_j(\beta)$, $j = 1, \ldots, n$.

- *initialization*
  Different strategies to obtain start values are possible:
  - Start with $\widehat{\beta}^{(0)}$, $\widehat{\eta}_j^{(0)}$ from the parametric (GLM) fit. Higher order polynomial terms in T may be included to allow for a nonlinear function $\widehat{\eta}_j^{(0)}$.
  - Alternatively, it is possible to use $\widehat{\beta}^{(0)} = 0$ and as in GLM $\widehat{\eta}_j^{(0)} = G^{-1}\{(Y_j + \overline{Y})/2\}$ (but $\widehat{\eta}_j^{(0)} = G^{-1}\{(Y_j + 0.5)/(m+1)\}$ for binomial responses).
  - Severini and Staniswalis (1994) propose to start with $\widehat{\beta}^{(0)} = 0$ and $\widehat{\eta}_j^{(0)} = G^{-1}(Y_j)$ (with an adjustment for binomial responses).

- *updating step for $\eta_j(\beta) = m_\beta(T_j)$*
  The function $\eta_j(\beta)$ is updated by

$$\widehat{\eta}_j^{(k+1)} = \widehat{\eta}_j^{(k)} - \frac{\sum_{i=1}^{n} \ell_i'(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k)})\,\mathcal{K}_{\mathbf{H}}(T_i - T_j)}{\sum_{i=1}^{n} \ell_i''(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k)})\,\mathcal{K}_{\mathbf{H}}(T_i - T_j)}.$$

- *updating step for β*
  The parameter β is updated by

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} - \mathcal{B}^{-1}\sum_{i=1}^{n} \ell_i'(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_i^{(k+1)})\widetilde{X}_i^{(k)}$$

  with a Hessian type matrix

$$\mathcal{B} = \sum_{i=1}^{n} \ell_i''(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_i^{(k+1)})\,\widetilde{X}_i^{(k)}\widetilde{X}_i^{(k)T}$$

  and

$$\widetilde{X}_j^{(k)} = X_j - \frac{\sum_{i=1}^{n} \ell_i''(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k+1)})\,\mathcal{K}_{\mathbf{H}}(T_i - T_j)\,X_i}{\sum_{i=1}^{n} \ell_i''(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k+1)})\,\mathcal{K}_{\mathbf{H}}(T_i - T_j)}.$$

As an alternative, the functions $\ell_i''(u)$ can be replaced by their expectations (w.r.t. to Y) to obtain a Fisher scoring type procedure.

2.2.2 Practical Application

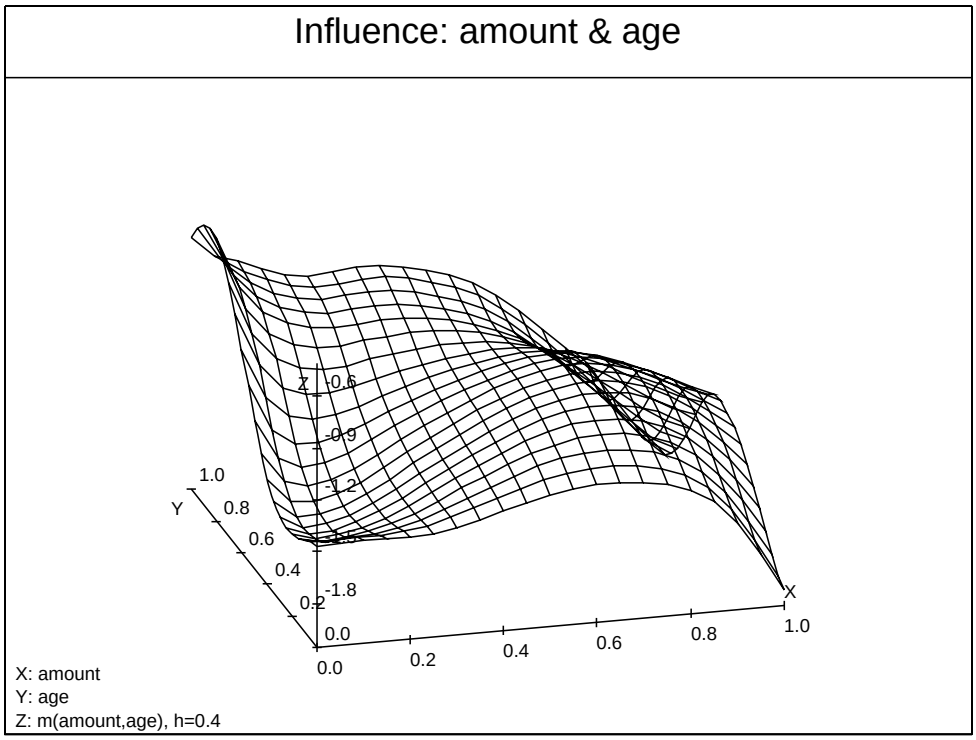
Let us illustrate the semiparametric estimation with the previously introduced credit scoring example, (Fahrmeir and Tutz, 1994; Fahrmeir and Hamerle, 1984). Recall that the data set consists of $n = 1000$ clients, among which 700 paid a credit back without problems and 300 did not. We define the binary variable Y with value 1 for those who paid back and 0 if not. The data set contains observations from three continuous variables (duration and amount of credit, age of client) and 17 discrete variables. The interest consists in finding how explanatory variables are related to credit worthiness.

	Coeff.	(t–value)	Coeff.	(t–value)	Coeff.	(t–value)
const.	-0.699	(-1.37)	-2.267	(-3.17)	–	–
duration	-2.179	(-3.63)	-2.242	(-3.73)	-2.527	(-4.23)
amount	-0.228	(-0.33)	5.652	(2.69)	–	–
amount squared	–	–	-5.470	(-2.97)	–	–
age	0.443	(1.04)	1.618	(1.06)	–	–
age squared	–	–	-1.232	(-0.78)	–	–
...	...	...	...	...	...	...
	Linear (logit)		Quadratic (logit)		Part. Linear	

Table 3: Logit coefficients and GPLM coefficients (t–values in parenthesis). $n = 1000$, $h_1 = h_2 = 0.4$ for GPLM. Credit data Fahrmeir and Hamerle (1984).

In the following statistical analysis we took logarithms of amount and age and transformed all explanatory variables linearly to the interval $[0, 1]$. A parametric logit model leads to the parameter estimates listed in Table 3. We omit the parameter estimates for the discrete explanatory variables. The influence of duration is highly significant. Amount and age have no significant coefficients if we include them linearly. We will see that the insignificant coefficients are a sign for a more complex structured influence, at least in the amount direction.

In a next step we fitted a generalized partially linear model according to the algorithm presented above. Here, the influence of amount and age has been fitted nonparametrically. Figure 8 shows the two–variate estimate $\widehat{m}$ (using a bandwidth $h = 0.4$ in both dimensions) in the upper panel. A scatterplot of amount versus age is given in the lower panel of Figure 8.



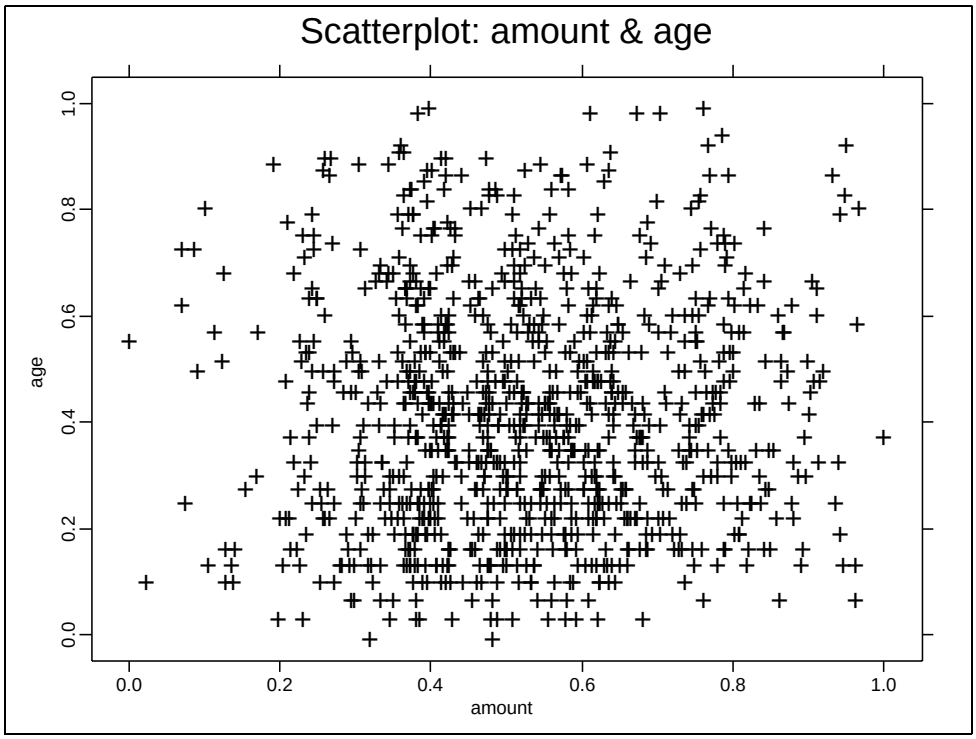


Figure 8: Two–dimensional nonparametric function of amount and age in GPLM (upper panel). $h_1 = h_2 = 0.4$. Scatterplot of of amount and age (lower panel). Credit data, $n = 1000$ Fahrmeir and Hamerle (1984).

It is difficult to check $\widehat{m}$ graphically for significant deviances from linearity. The high values of $\widehat{m}$ are caused by only a few observations (as can be seen from the scatterplot). For a closer inspection of $\widehat{m}$ Figure 9 shows a contour plot of $\widehat{m}$. It reveals that we have more nonlinear influence in the amount than in the age direction. For comparison we also fitted a GPLM where only amount is included in the nonparametric way. Figure 10 show the resulting nonparametric estimates for different bandwidth ($h = 0.2, \ldots, 0.5$).

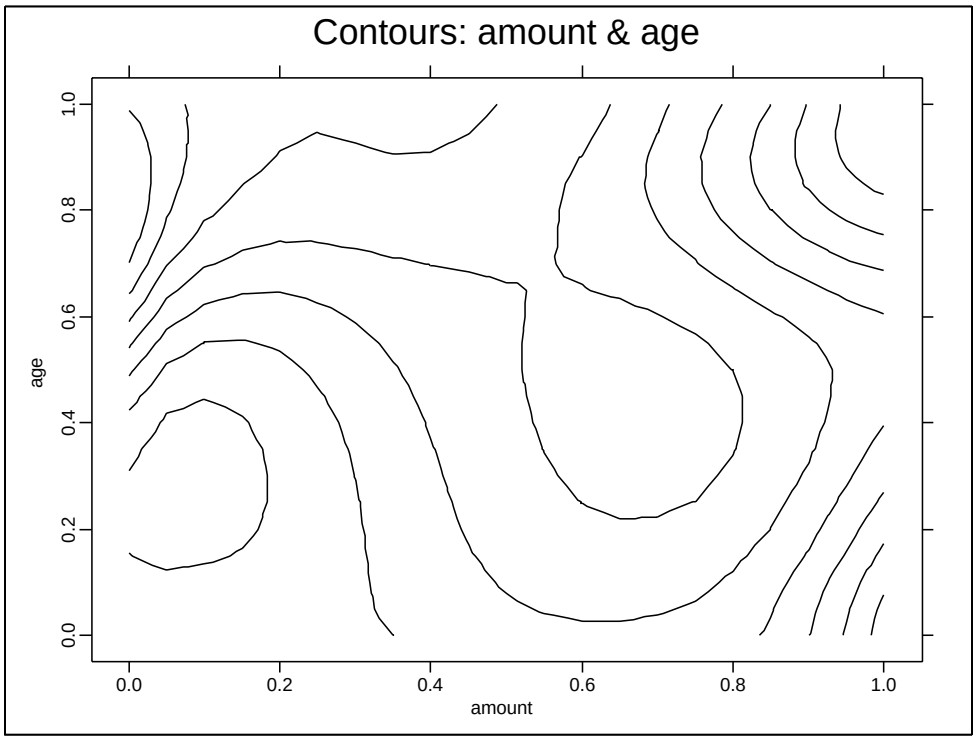


Figure 9: Contours for function of amount and age in GPLM. $h_1 = h_2 = 0.4$. Credit data, $n = 1000$ Fahrmeir and Hamerle (1984).

Since the question of an optimal bandwidth selection is still open for generalized partial linear models, we have carried out the analysis for different bandwidths. The nonparametric estimates $\widehat{m}$ for the different bandwidths are obviously nonlinear functions. However, it is difficult to judge whether a nonparametric estimate gives a significant improvement. In general, it cannot be excluded that the difference between the nonparametric and the linear fit may be caused by boundary and bias problems of $\widehat{m}$. Additionally, some of the other (discrete) covariables have a quite dominant influence on credit worthiness.

Härdle, Mammen and Müller (1996) proposed a procedure for testing GLM versus GPLM. We applied this test using and computed critical values from the approximative normal distribution. Table 4 shows the observed significance levels for rejection. The decision of the test depends obviously on the bandwidth. As Härdle, Mammen and Müller (1996) point out, this is due to a slow convergence of the test statistic towards its

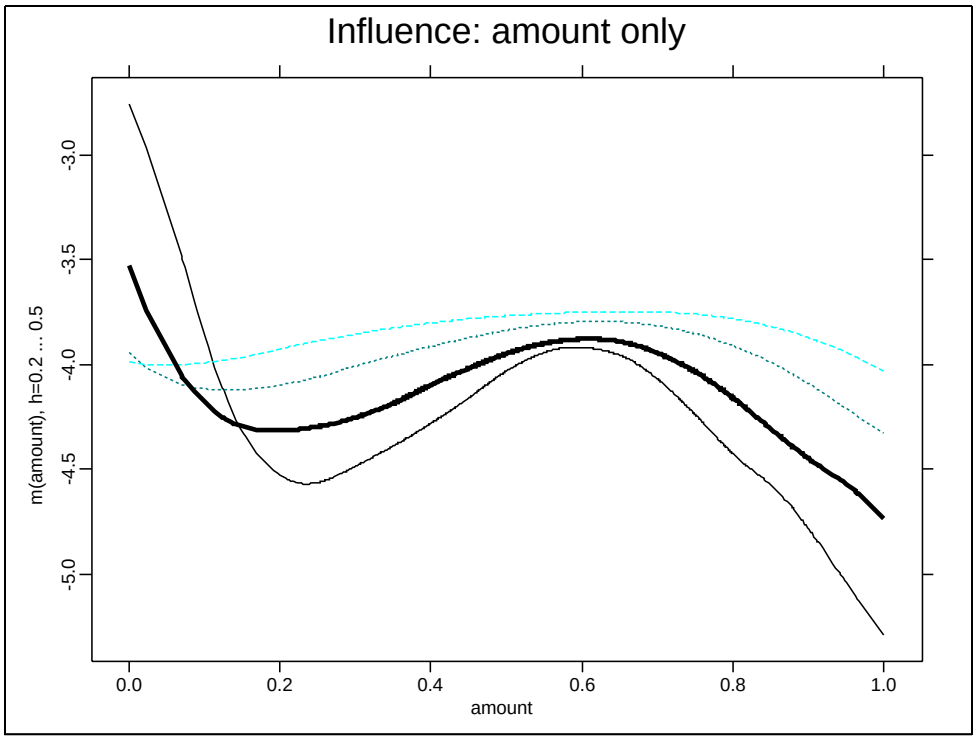


Figure 10: Nonparametric function of amount only in GPLM. $h = 0.3$ (thick line), $h = 0.2$, $h = 0.4$, $h = 0.5$. Credit data, $n = 1000$ Fahrmeir and Hamerle (1984).

limiting normal distribution and can be "repaired" by applying a bootstrap version of the test. We omit the details here.

h	0.1	0.2	0.3	0.4	0.5	0.6
amount only	<0.01	<0.01	<0.01	0.45	0.56	–
amount and age	–	<0.01	<0.01	0.08	0.26	0.54

Table 4: Observed significance levels for linearity test. Credit data, Fahrmeir and Hamerle (1984).

We see from Table 4 that linearity is clearly rejected for bandwidths 0.1 to 0.3 for the univariate nonparametric component (amount only) and 0.2 to 0.4 for the bivariate nonparametric component (amount and age). Including only age with an nonlinear influence shows no significant test result. This is in accordance with the parametric inclusion of quadratic terms in age. Obviously, the joint nonlinear effect of both amount and age is mainly determined by amount.

For higher dimensions in T the possible nonlinearities in (46) cannot anymore be graphically displayed and face the above mentioned problems (interpretability). An additive

structured partial linear index may be considered. This is considered in Hastie and Tibshirani (1990) on basis of the backfitting algorithm. A variant based on the integration method introduced by Linton and Nielsen (1994) is currently under development, see Härdle, Huet, Mammen and Sperlich (1996).

References

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1995). Generalized partially linear single–index models, *Discussion Paper 9506*, Institut de Statistique, Université Catholique, Louvain–La–Neuve.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numer. Math.* **31**: 377–403.

Fahrmeir, L. and Hamerle, A. (1984). *Multivariate Statistische Verfahren*, De Gruyter, Berlin.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M. and Engel, J. (1993). Local polynomial fitting: A standard for nonparametric regression, *Discussion Paper 9315*, Institut de Statistique, Université Catholique, Louvain–La–Neuve.

Fan, J. and Gijbels, I. (1995). *Local Polynomial Modeling and Its Application — Theory and Methodologies*, Chapman and Hall, New York.

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics* **3**(1): 35–56.

Fan, J. and Müller, M. (1995). Density and regression smoothing, *in* W. Härdle, S. Klinke and B. A. Turlach (eds), *XploRe – an interactive statistical computing environment*, Springer, pp. 77–99.

Härdle, W. (1990). *Applied Nonparametric Regression*, Econometric Society Monographs No. 19, Cambridge University Press.

Härdle, W. (1991). *Smoothing Techniques, With Implementations in S*, Springer, New York.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression estimators from their optimum?, *Journal of the American Statistical Association* **83**: 86–97.

Härdle, W., Hall, P. and Marron, J. S. (1992). Regression smoothing estimators that are not far from their optimum, *Journal of the American Statistical Association* **87**: 227–233.

Härdle, W., Huet, S., Mammen, E. and Sperlich, S. (1996). Semiparametric additive indices for binary response, *Technical report*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Härdle, W., Mammen, E. and Müller, M. (1996). Testing parametric versus semiparametric modelling in generalized linear models, *SFB 373 Discussion Paper 28*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Härdle, W. and Scott, D. (1992). Smoothing in by weighted averaging using rounded points, *Computational Statistics* **7**: 97–128.

Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives, *Journal of the American Statistical Association* **84**: 986–995.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Horowitz, J. L. (1993). Semiparametric and nonparametric estimation of quantal response models, *in* G. S. Madala, C. R. Rao and H. D. Vinod (eds), *Handbook of Statistics*, Elsevier Science Publishers, pp. 45–72.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single index models with discrete covariates, *Journal of the American Statistical Association*. to appear.

Hunsberger, S. (1994). Semiparametric regression in likelihood–based models, *Journal of the American Statistical Association* **89**: 1354–1365.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single–index models, *Journal of Econometrics* **58**: 71–120.

Klein, R. and Spady, R. (1993). An efficient semiparametric estimator for binary response models, *Econometrica* **61**: 387–421.

Korostelev, A. and Müller, M. (1995). Single index models with mixed discrete-continuous explanatory variables, *Discussion Paper 26*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Linton, O. and Nielsen, J. P. (1994). A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika*. in press.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*, Econometric Society Monographs No. 4, Cambridge University Press.

Marron, J. S. and Nolan, D. (1988). Canonical kernels for density estimation, *Statistics & Probability Letters* **7**(3): 195–199.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, 2 edn, Chapman and Hall, London.

Müller, H.-G. (ed.) (1988). *Nonparametric Regression Analysis of Longitudinal Data*, Springer, Berlin.

Newey, W. and Stoker, T. (1993). Efficiency of weighted average derivative estimators and index models, *Econometrica* **5**: 1199–1223.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients, *Econometrica* **57**(6): 1403–1430.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression, *Annals of Statistics* **22**(3): 1346–1370.

Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994). Cross–validation of multivariate densities, *Journal of the American Statistical Association* **89**(427): 807–817.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York, Chichester.

Scott, D. and Wand, M. (1991). Feasibility of multivariate density estimates, *Biometrika* **78**: 197–205.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**: 501–511.

Severini, T. A. and Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models, *Annals of Statistics* **20**: 1768–1802.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Stoker, T. M. (1991). Equivalence of direct, indirect and slope estimators of average derivatives, *in* W. A. Barnett, J. Powell and G. Tauchen (eds), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Proceedings of the fifth international symposium in economic theory and econometrics, Cambridge University Press.

Wand, M. P. (1994). Fast computation of multivariate kernel estimators, *Journal of Computational and Graphical Statistics* **3**(4): 433–445.

Wand, M. P. and Jones, M. C. (1994). Multivariate plug–in bandwidth selection, *Computational Statistics* **9**: 97–911.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Weisberg, S. and Welsh, A. H. (1994). Adapting for the missing link, *Annals of Statistics* **22**: 1674–1700.